3 Marker Regression Analysis

- marker regression in a backcross
- marker regression in a F₂ intercross
- marker regression by linear regression
- LOD scores
- LOD thresholds
- advantages and disadvantages

3.1 marker regression in a backcross

- consider backcross of P₁ to F₁=P₁xP₂
  - sample size \( n \approx 100-500 \) individuals
  - collection of \( m \approx 75-300 \) markers
    - not necessarily arranged as a linkage map
- goal: identify markers linked to a QTL
  - consider each marker individually
  - split individuals into 2 groups by marker genotype
- examine/test for difference between groups
  - plot data
  - hypothesis test of no QLT vs. QTL linked to marker
Sugiyama et al. (2001)

- salt-induced hypertension
  - 250 mice (B6 x A) x B6 backcross
  - C57BL/6J (A) and A/J (a) strains
- genotyped at 173 markers
  - 19 mouse chromosomes (autosomes)
  - selective genotyping of 92 mice on most (later)
- hypothesize one QTL in genome
  - consider markers one at a time
  - QTL exactly at marker or just linked?

phenotype split by genotype

- jittered dot plots
- confidence intervals: mean ± 2SE
- D4Mit214: B6 (AA genotype) has more hypertension
- D12Mit20: no apparent difference

considerable spread
environmental?
more QTL?
clinically substantial?
estimating genotype values & SDs

- \( G_{AA}, G_{Aa} \) phenotype means for AA, Aa
  - estimated by within-group sample averages
- common standard deviation (SD) of \( \sigma \)
  - weighted average of within-group SDs
- form two-sample \( t \) test statistic
  - null hypothesis: no QTL, \( G_{AA} = G_{Aa} \)
  - reject for large values of \(|t|\)
- cautions/interpretation
  - how to convert to LOD score?
  - how to account for multiple testing across \( m \) markers?

\[
\hat{G}_{AA} = \sum_i \left( Y_i 1(X_{ij} = AA) \right) / n_{AA} \\
s_{AA}^2 = \sum_i \left( (Y_i - \hat{G}_{AA})^2 1(X_{ij} = Aa) \right) / n_{AA} \\
SD = \hat{\sigma}_{pool} = \sqrt{ \frac{(n_{AA} - 1)s_{AA}^2 + (n_{Aa} - 1)s_{Aa}^2}{n_{AA} + n_{Aa} - 2} } \\
t = \frac{\hat{G}_{AA} - \hat{G}_{Aa}}{\hat{\sigma}_{pool} \sqrt{1/n_{AA} + 1/n_{Aa}}} 
\]
data analysis at two markers

- D4Mit214: \( n_{AA} = 130, n_{Aa} = 120 \)
  \[ G_{AA} = 104.4, G_{aa} = 98.6, \text{ SD } = 7.92, t = 5.78 \]
- D12Mit20: \( n_{AA} = 124, n_{Aa} = 126 \)
  \[ G_{AA} = 101.5, G_{aa} = 101.7, \text{ SD } = 8.44, t = -0.25 \]

high \( t \) statistics near D4Mit214—why?
why do some nearby markers have small \( t \) values?
assume only 1 QTL in region ...

\[ G_{AA} = (1 - r) \mu_{AA} + r \mu_{Aa} = \mu_{AA} - r(\mu_{AA} - \mu_{Aa}) = \mu_{AA} - r \Delta \]
\[ G_{aa} = (1 - r) \mu_{Aa} + r \mu_{AA} = \mu_{Aa} + r(\mu_{AA} - \mu_{Aa}) = \mu_{Aa} + r \Delta \]

actual vs. apparent QTL effect

- QTL linked to marker
  - recombination \( r \) between marker and QTL
  - not all \( n_{AA} \) have AA genotype at QTL
- means at marker \((G_{AA}, G_{Aa})\) vs. QTL \((\mu_{AA}, \mu_{Aa})\)
  \[ G_{AA} = (1 - r) \mu_{AA} + r \mu_{Aa} = \mu_{AA} - r(\mu_{AA} - \mu_{Aa}) = \mu_{AA} - r \Delta \]
  \[ G_{aa} = (1 - r) \mu_{Aa} + r \mu_{AA} = \mu_{Aa} + r(\mu_{AA} - \mu_{Aa}) = \mu_{Aa} + r \Delta \]
- apparent effect at marker (attenuated by \( r \))
  \[ G_{AA} - G_{Aa} = (\mu_{AA} - r \Delta) - (\mu_{Aa} + r \Delta) = (1 - 2r) \Delta \]
  \[ G_{AA} - G_{Aa} = \Delta \text{ if } r = 0, G_{AA} - G_{Aa} = 0 \text{ if } r = 0.5 \]
3.2 marker regression in F2 intercross

- 3 genotypes, split individuals into 3 groups
  - D1Mit100 shows higher mean for AA
  - D2Mit101 shows no apparent differences

hypothesis tests for F2

- all means identical at marker
  - null hypothesis: no QTL, $G_{AA} = G_{Aa} = G_{aa}$
  - alternative hypothesis: marker linked to QTL
  - assume constant variance

- analysis of variance
  - use $F$ statistics in place of $t$ statistics
  - reject for large $F$ in favor of linked QTL
3.3 marker regression by linear regression

- **what?**
  - recode marker genotype as numeric value(s)
  - set up regression to capture group means
  - test regression slopes = test of group means

- **why?**
  - always nice to have another perspective
  - can extend idea to multiple QTL
  - can help sort out genetic architecture details

- **how?**
  - see usual coding on next slide
  - other codings are preferred for multiple QTL (later)

### a regression recoding

<table>
<thead>
<tr>
<th>recode</th>
<th>genotypes</th>
<th>use for</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{ij}$</td>
<td>AA</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td>Aa</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>aa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aa</td>
<td>+1</td>
</tr>
<tr>
<td>$A_{ij}$</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>aa</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{ij}$</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
Y_i = \mu + \beta X_{ij} + e_i \quad \text{backcross}
\]

\[
Y_i = \mu + \alpha A_{ij} + \delta D_{ij} + e_i \quad \text{F2 intercross}
\]
3.4 LOD scores

- LOD scores and F statistics
  - both test null hypothesis of no QTL vs 1 QTL
- F statistics
  - evaluated using F tables (model and error d.f.)
  - based on quadratic forms, linear models
- LOD scores
  - evaluated using chi-square tables (model d.f.)
  - based on large-sample likelihood principle
  - can handle more complicated model forms
  - LOD is approximately proportional to F statistic

LOD score for 1 QTL, F2

- compare null to QTL model
- QTL at marker $j$
- $f$ = normal density function

$$L_0(\hat{\mu}, \hat{s}^2 | Y) = \prod_i f(Y_i | \hat{\mu}, \hat{s}^2)$$

$$L(\hat{G}, \hat{\sigma}_{pool}^2 | Y, X) = \prod_i f(Y_i | \hat{G}_{X,j}, \hat{\sigma}_{pool}^2)$$

$$LOD = \log_{10} \left( \frac{L(\hat{G}, \hat{\sigma}_{pool}^2 | Y, X)}{L_0(\hat{\mu}, \hat{s}^2 | Y)} \right)$$
3.5 LOD thresholds

• how large does a LOD have to be?
  – evaluate LOD under null of no QTL
    • recall chi-square distribution
    – but adjust for many, many tests
• want genome-wide threshold
  – has to be bigger than for a single test
  – depends on genome size, cross, number of markers, missing data, phenotype distribution

genome-wide threshold

• dashed = 1 marker
• solid = genome-wide
• backcross (idealized)
• often use 95%-ile

• how to evaluate genome-wide threshold?
  – what is maximum LOD over entire genome under null?
  – theory, simulation, or permutation
  – permutation is recommended
genome-wide permutation

- permute phenotypes
  - 1000 times, say
  - random shuffle
  - same genotype data
- compute max LOD
- draw histogram
- find 95%-ile
  - is max LOD from data above this value?

3.6 advantages & disadvantages

- advantages
  - simple: test all markers with $t$, $F$, or LOD
  - can use standard statistical software
    - easy to incorporate covariates, interactions, design
  - no need for genetic map
- disadvantages
  - discard individuals with missing data at marker
  - cannot inspect positions between markers
  - recombination rate and QTL effect are confounded
  - considers only 1 QTL at a time
    - can use multiple regression on multiple markers (dense map)
    - but missing genotype problem is compounded